YOLOv11-Based Detection of Blue Whale and Fin Whale Vocalizations via Spectrogram Object Detection

BioDCASE 2026 Challenge – Task 2

Marie Vogler
University of Hamburg
vogler.marie@outlook.de
Alex Savchik
Alfred Wegener Institute of Polar and Marine Research
alexey.savchik@awi.de

Abstract

In this solution, we update the baseline (YOLOv11s) model[1] with an extra per-class confidence threshold optimisation step.

The baseline frames the detection of blue whale and fin whale vocalizations as an object detection task on spectrogram images. The audio is segmented into overlapping 30-second chunks, converted to spectrogram images, and fed to the YOLOv11s model. Three merged call classes are targeted: Antarctic blue whale A/B/Z calls (`bmabz`), D-type calls from both species (`d`), and fin whale 20 Hz calls (`bp`).

The additional per-class confidence threshold calibration step, performed offline on a held-out validation set, compensates for large differences in class-wise detector confidence distributions. Per-class confidence threshold optimisation on the validation set yields F1 scores of 0.252, 0.469, and 0.431 for `bmabz`, `d`, and `bp`, respectively (F1 = 0.384 on average).

1 Introduction

Passive acoustic monitoring (PAM) is a key tool for studying the distribution and behaviour of large baleen whales such as the Antarctic blue whale (*Balaenoptera musculus intermedia*) and the fin whale (*Balaenoptera physalus*). Continuous hydrophone recordings can span months or years, making manual annotation infeasible and automated detection indispensable.

BioDCASE 2026 Task 2 challenges participants to localise whale call events in time and frequency given raw waveform recordings from multiple deployments spanning different years and geographic locations. The evaluation metric is macro-averaged F1 computed with a temporal IoU threshold of 0.3.

Our approach treats each 30-second audio segment as an image and applies the YOLOv11s object detector [1] to predict the onset time, offset time, and frequency extent of each call directly from a magnitude spectrogram image. This enables the system to benefit from strong ImageNet pre-training while being fine-tuned end-to-end on challenge data.

[1] https://github.com/marinebioCASE/task2_2025_2026/tree/main/baselines/yolo

2 Preprocessing

2.1 Audio segmentation

Each WAV file (sampled at 250 Hz) is segmented into overlapping 30-second chunks. Consecutive chunks are offset by 15 seconds, yielding a 50% overlap. This ensures that vocalization events spanning chunk boundaries appear fully contained within at least one chunk.

2.2 Bandpass filtering

Before spectrogram computation, each chunk is filtered with a Butterworth bandpass filter (passband 5–124 Hz, filter order 20) implemented as a cascade of second-order sections via `scipy.signal.iirfilter`. The filter removes very low-frequency environmental noise (ocean swell, shipping) and suppresses energy above the Nyquist frequency of the target calls.

2.3 Spectrogram generation

A short-time Fourier transform (STFT) spectrogram is computed using `scipy.signal.spectrogram` with the following parameters:

- Sampling rate: 250 Hz
- Window function: Hann, length 256 samples (≈1.02 s)
- Hop size: 20 samples (≈0.08 s)
- FFT size: 512 points
- Output mode: linear magnitude (no log scaling, no colour map)

The resulting spectrogram covers 0–125 Hz with a frequency resolution of approximately 0.49 Hz and a temporal resolution of 0.08 s per frame.

2.4 Normalisation and image encoding

The magnitude spectrogram S is contrast-inverted ($S \leftarrow 1 - S$) so that high-energy vocalization events become darker than background. The inverted spectrogram is then linearly

rescaled to $[0, 255]$ using the 98th percentile of the pixel values as the upper bound; values exceeding this percentile are clipped. The result is saved as an 8-bit greyscale PNG image with the frequency axis vertically flipped so that low frequencies appear at the bottom. YOLO internally resizes all images to 640×640 pixels via bilinear interpolation.

3 Dataset Preparation

3.1 Class merging

The challenge provides seven fine-grained call-type annotations. To reduce inter-class confusion and increase per-class training sample counts, these are merged into three YOLO target classes (Table 1). Class `bmabz` groups Antarctic blue whale A, B, and Z calls, which share similar tonal, patterned structures at low frequencies. Class `d` combines D-type downsweep calls produced by both species. Class `bp` covers the two variants of the pygmy blue whale 20 Hz call.

Table 1: Class merging scheme.

YOLO ID	Label	Source annotations
0	`bmabz`	bma, bmb, bmz
1	`d`	bmd, bpd
2	`bp`	bp20, bp20plus

3.2 Background sampling

The majority of 30-second chunks contain no annotated whale calls. To avoid severe class imbalance during training, background (no-call) chunks are sampled such that the number of selected background chunks matches the number of annotation-containing chunks within each hydrophone deployment. This yields a balanced training set with approximately equal proportions of positive and background examples per deployment.

4 Model Architecture and Training

4.1 Architecture

YOLOv11s (the small variant of the Ultralytics YOLO v11 family) is used as the detection model. The architecture consists of a CSPDarkNet-inspired backbone (layers 0–10), a Path Aggregation Feature Pyramid Network neck (layers 11–22), and a decoupled detection head (layer 23). The model contains approximately 9.4 million parameters and produces a checkpoint of approximately 19 MB.

4.2 Training configuration

The model is initialised from the ImageNet pre-trained weights `yolo11s.pt` provided by Ultralytics. The challenge development dataset is partitioned into a training split and a held-out validation split; the model is trained exclusively on the training split. Key hyperparameters are listed in Table 2.

Table 2: Training hyperparameters.

Hyperparameter	Value
Epochs	20
Batch size	32
Image size	640×640
Training IoU	0.3
Random seed	24
Devices	2 GPUs

All geometric and photometric augmentation options available in the Ultralytics framework are disabled (flipping, scaling, HSV jitter, mosaic, MixUp, copy-paste, random erasing). This prevents the introduction of frequency patterns or temporal distortions that are physically unrealistic for spectrogram images. The best model checkpoint, selected according to validation mAP50-95 over the course of training, is used for all subsequent inference.

During inference, class-agnostic non-maximum suppression (NMS) is applied so that overlapping bounding boxes from different call classes can suppress each other.

5 Confidence Threshold Optimisation

5.1 Motivation

A single global confidence threshold is suboptimal when the three call classes differ substantially in training density and acoustic distinctiveness. For example, with a low global threshold the `d` class accumulates a large number of false positives, whereas a high global threshold causes excessive misses for less-prominent classes. We therefore determine a separate optimal threshold for each class.

5.2 Threshold sweep procedure

Predictions are generated on the held-out validation set at a very low base confidence threshold of 0.001, retaining all candidate detections. A threshold sweep is then performed over the collected prediction file: for each candidate threshold τ, detections with confidence $\geq \tau$ are retained for each class independently, matched against the ground-truth annotations using an IoU threshold of 0.3, and the per-class F1 score is computed. The threshold that maximises F1 for each class is selected independently.

5.3 Resulting thresholds

Table 3 lists the selected per-class thresholds and the corresponding per-class F1 scores on the validation set. During inference the model is run at $\tau_{\min} = \min(0.015, 0.25, 0.003) = 0.003$ so that all candidate detections are available; detections are then filtered per class using their respective threshold.

Table 3: Per-class thresholds and validation F1 scores (IoU = 0.3).

Class	Threshold τ	Validation F1
bmabz	0.015	0.252
d	0.250	0.469
bp	0.003	0.431
Macro	–	0.384

6 Results

Table 4 reports the macro-averaged F1, precision, and recall on the three development-set subsets, evaluated at an IoU matching threshold of 0.3. All values represent macro averages over the three call classes (bmabz, d, bp). The challenge evaluation set consists of four previously unseen hydrophone deployments: DDU2019, DDU2021, Kerguelen2019, and Kerguelen2020.

Table 4: Development set results (macro over 3 classes, IoU = 0.3).

Dataset	Macro F1	Macro P	Macro R
Casey2017	0.21	0.62	0.15
Kerguelen2014	0.17	0.76	0.11
Kerguelen2015	0.21	0.74	0.13
Overall	0.20	0.77	0.13

The d class consistently achieves the highest per-class F1 across all three subsets (0.50, 0.32, 0.43). Both bmabz and bp exhibit high precision but very low recall (5–7% and 4%, respectively), indicating that the model is conservative in detecting these classes. This likely reflects the limited acoustic diversity seen during training: Antarctic blue whale A/B/Z calls and pygmy blue whale 20 Hz calls may vary substantially in time–frequency structure across deployments, making it difficult to generalise from the development recordings. The overall macro precision of 0.77 suggests that the detections that *are* made are largely correct.

7 Conclusion

We applied YOLOv11s as a spectrogram object detector for whale call localisation and demonstrated that a per-class confidence threshold calibration step markedly improves detection performance compared to a single global threshold, especially for the d class. The system is conceptually simple, leverages strong ImageNet pre-training, and requires no handcrafted acoustic features.

Future work should explore log-frequency spectrograms (to emphasise relative frequency relationships), domain-specific data augmentation (e.g., SpecAugment or frequency masking), a finer-grained class taxonomy, and threshold calibration on a larger or more representative validation set to improve cross-deployment generalisation.

References

[1] Ultralytics, *YOLO11: Real-Time Object Detection*, https://docs.ultralytics.com/models/yolo11/, 2024.